NIH Security Best Practices for Controlled-Access Data Subject to the NIH Genomic Data Sharing (GDS) Policy

Updated: 09 MAR 2015

Introduction

This document is intended for officials at academic institutions and scientific organizations whose investigators are granted access under the NIH Genomic Data Sharing (GDS) Policy to controlled-access human genomic and phenotypic data that are maintained in NIH-designated data repositories.[1] It provides an outline of the NIH's expectations for the management and protection of NIH controlled access data transferred to and maintained by institutions whether in their own institutional data storage systems or in cloud computing systems.[2] Although controlled-access data do not contain direct identifiers, the data are sensitive and must be protected. The principles governing access and use of such data are outlined in the GDS Policy and individual Data Use Certification (DUC) Agreements that investigators submit as part of the process of requesting access to controlled access data. This process is intended to ensure that NIH controlled-access genomic and phenotypic data are kept secure and no one other than users approved by NIH is able to access the data.

The information contained in this document is targeted at two distinct audiences: scientific professionals including institutional signing officials and investigators that will use the data, and information technology professionals, including Chief Information Officers (CIOs), Information Systems Security Officer (ISSOs) and operations staff working for both central IT organizations and embedded within research groups. Accordingly this document is split into two main sections focused on each of these groups.

Information for Scientific and Administrative Staff

General Considerations

Under the GDS Policy, the recipient institution is ultimately responsible for maintaining the confidentiality, integrity and availability of the data to which it is entrusted by the NIH. Failure to provide appropriate controls can subject investigators or institutions to sanctions defined by the GDS Policy as well as significantly erode public confidence in the ability of NIH and its grantees to carry out research using sensitive information. It is therefore essential that all recipients of controlled access data understand their responsibilities for ensuring appropriate information security controls and that the work with their IT organizations to effectively implement those responsibilities.

---

[1] dbGaP and the Sequence Read Archive are examples of NIH-designated data repositories. NIH has also established Trusted Partnerships with several institutions to serve as NIH-designated data repositories.

[2] The National Institute of Standards and Technology defines cloud computing as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. See: http://csrc.nist.gov/publications/nistpubs/800-145/SP800145.pdf.

The NIH provides this best practice document so that institutions can obtain an understanding of the types of information security practices that they should be enacting. However, this best practice document is not a substitute for a more formal security plan that is devised for the specific local or cloud configuration chosen by the investigators and institution.

The NIH strongly recommends that investigators consult with institutional IT leaders, including the Chief Information Officer (CIO) and the institutional Information Systems Security Officer (ISSO) or equivalents to develop the formal information security plan prior to receipt of controlled access data from the NIH, and ***institutional signing officials should validate that an appropriate security plan is in place prior to accepting liability for data loss or breach on behalf of the institution***. This document provides an overview of security principles for data, access, and physical security to ensure confidentiality, privacy, and accessibility of data. This is a minimum set of requirements; additional restrictions may be needed by your institution and should be guided by the knowledge of the user community at your institution as well as your institution's IT requirements and policies.

The single most important element (regardless of type of infrastructure) for maintaining the security of NIH controlled access data is to design security into the chosen environment before the data is transferred rather than attempting to add security controls to an environment after the data has been downloaded. Security controls should be on by default; investigators and users should not have to perform any active action to turn them on. To use an analogy, doors should be locked by default rather than need to be actively locked by someone. A corollary is that all users and support staff associated with the project need to have an information security mindset going into the project, and all must be aware that public support for the collection and dissemination of these types of data are their individual responsibilities, and it is essential that all staff members that will interact with the data or the systems that maintain the data have appropriate information security training. This is particularly true for groups that wish to use cloud computing, and in these cases, NIH recommends additional training to inform staff of the special risks that the use of such infrastructure entails.

Part of having an information security mindset is being aware of the multiple dimensions of ***access control*** and ***accountability*** at all times. This means ensuring that passwords and/or access devices (smart cards, soft or physical tokens, etc.) are physically safe, strong and not shared with anyone and that data is both physically and logically (i.e. electronically) secure. Particular care must be taken with copies of data on portable electronic media and devices (i.e. laptops, tablets, USB thumb drives, tapes, etc.). Generally speaking, users should avoid putting controlled access data on such devices wherever possible. If it is necessary, such devices must be encrypted and should be treated as if they are cash, with appropriate physical and electronic controls, including remote wipe capability wherever possible. In addition, please remember that collaborators at different institutions must file a separate data access request ***even if they are working on the same project.***

Finally, remember that data downloaded from NIH-designated data repositories must be destroyed if they are no longer needed or used, or if the project is to be terminated and closed-out in the dbGaP Authorized Access System. Investigators may retain only encrypted copies of the minimum data necessary at their institution to comply with institutional scientific data retention policy and any data stored on temporary backup media as are required to maintain the integrity of the general institutional data protection (i.e. backup) program.

Additional Information Related to the Use of Cloud Computing

Cloud computing, as defined by the National Institute for Standards and Technology (NIST), is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or cloud service provider interaction. In contrast to traditional computing on local servers and hardware, cloud computing often entails the transfer and storage of controlled-access data on systems managed by a third party. Cloud computing offers a number of advantages for authorized investigators but also requires additional security considerations.

Most of the recommendations described above apply to cloud computing; indeed, the primary difference is that while information security in cloud environments is still the responsibility of the institution, ***the implementation of that security is shared between the institution and the cloud service provider.*** Thus, it is essential that institutions validate that they are partnering with a reputable cloud service provider. Institutions should ensure that they understand the security policies and practices utilized and recommended by their cloud service provider of choice, and may wish to obtain third party reviews or audits from the cloud service provider. Institutions should utilize these best practices, work with their cloud service provider to understand and implement the best practices associated with their specific environment and ensure that the cloud service provider can meet institutional information security requirements. Because the use of cloud computing has the potential for being higher risk than using local infrastructure, the NIH strongly recommends that you consult with your institutional CIO, ISSO and IT staff to ensure that an appropriate security plan is developed and that necessary technical, training and policy controls are in place before data is migrated to cloud environments. Remember – ***you and your institution are accountable for ensuring the security of this data, not the cloud service provider.***

Information for IT Professionals

Local Infrastructure Guidance

General Information Security Guidelines

- When using local infrastructure, make sure these files are never exposed to the Internet with the exception of such connections as are required to download data from source repositories. Infrastructure should be behind local and/or institutional firewalls that block access from outside of the institution. For cloud infrastructure, investigators must restrict external access to instances and storage under the investigator's control (see section on cloud computing for more details).

- Data must never be posted on servers in any fashion that will make them publically accessible, such as an investigator's (or institution's) website, because the files can be "discovered" by Internet search engines, e.g., Google, Bing.

- Institutions must not set up web or other electronic services that host data publicly, or that provide access to other individuals that are not listed on the Data Use Request even if those individuals have access to the same dbGaP data. Providing such access requires that an organization be an NIH Trusted Partner, with different requirements above and beyond those required for access to NIH controlled data.

- Utilize strong authentication technology for access control. Two factor authentication technologies (smart cards, hard or soft token, etc.) are preferred. When using single factor passwords, set policies that mandate the following requirements:

    - Minimum length of 12 characters
    - Does not contain user names, real names or company names o Does not contain a complete dictionary word
    - Contains characters from each of the following groups: lowercase letters, uppercase letters, numerals, and special characters
    - Passwords should expire every 120 days or at the rate required by institutional policies, whichever is more frequent.

- Avoid allowing users to place controlled access data on mobile devices (e.g. laptops, smartphones, tablets, mp3 players) or removable media such as USB thumb drives (except where such media are used as backups and follow appropriate physical security controls). If data must be placed on mobile devices, it must be encrypted. NIH recommends the use of NIST validated encryption technologies.

- Keep all software patches up-to-date.

Physical Security Guidelines

- Data that are in hard copy or reside on portable media, e.g., on a USB stick, CD, flash drive or laptop should be treated as though it were cash, with appropriate controls in place. Such media must be encrypted and stored in a secured in a locked facility with access granted to the minimum number of individuals required to efficiently carry out research.

- Restrict physical access to all servers, network hardware, storage arrays, firewalls and backup media only to those that are required for efficient operations.

- Log access to secure facilities, ideally with electronic authentication.

Controls for Servers

- Keep servers from being accessible directly from the Internet, (i.e. must be behind a firewall or not connected to a larger network) and disable unnecessary services. It is better to begin with a server image that disables all non-essential services and restore those that are needed than to start with a full-featured image and disable unnecessary services.

- Enforce principle of Least Privilege to ensure that individuals and/or processes grant only the rights and permissions to perform their assigned tasks and functions, but no more.

- Secure controlled-access genomic and phenotypic data on the systems from other users (restrict directory permissions to only the owner and group) and if exported via file sharing, ensure limited access to remote systems.

- If accessing systems remotely, use encrypted data access (such as Secure Shell (SSH) or Virtual Private Network (VPN)). It is preferred to use a tool such as Remote Desktop (RDP), X-windows or Virtual Network Computing (VNC) that does not permit copying of data and provides "View only" support.

- If data is used on multiple systems (such as a compute cluster), ensure that data access policies are retained throughout the processing of the data on all the other systems. If data is cached on local systems, directory protection must be kept, and data must be removed when processing is complete. Requesting investigators must meet the spirit and intent of these protection requirements to ensure a secure environment 24 hours a day for the period of the agreement.

Source Data and Control of Copies of Data

- Approved users must retain the original version of the encrypted data, track all copies or extracts and ensure that the information is not divulged to anyone except authorized staff members at the institution. NIH therefore recommends ensuring careful control of physical copies of data and providing appropriate logging on machines where such data is resident.

- As collaborating investigators from other institutions must submit an independent DAR and be approved by NIH to access to the data, restrict outbound access from devices that host controlled access data.

Destruction of Data

- Data downloaded from NIH-designated data repositories must be destroyed if they are no longer needed or used, or if the project is to be terminated and closed-out in the dbGaP Authorized Access System. Delete all data for the project from storage, virtual and physical machines, databases, and random access archives (i.e., archival technology that allows for deletion of specified records within the context of media containing multiple records).

- Investigators and Institutions may retain only encrypted copies of the minimum data necessary at their institution to comply with institutional scientific data retention policy and any data stored on temporary backup media as are required to maintain the integrity of the institution's data protection program. Ideally, the data will exist on backup media that is not used by other projects and can therefore be destroyed or erased without impacting other users/tenants. If retaining the data on separate backup media is not possible, as will be the case with many users, the media may be retained for the standard media retention period but may not be recovered for any purpose

without a new Data Access Request approved by the NIH. Retained data should be deleted at the appropriate time, according to institutional policies.

- Shred hard copies and CD ROMs or other non-reusable physical media.

- Delete electronic files securely. For personal computers, the minimum would involve deleting files and emptying the recycle bin or equivalent with equivalent procedures for servers. Optimally, use a secure method that performs a delete and overwrite of the physical media that was used to store the files.

- Ensure that backups are reused (data deleted) and any archive copies are also destroyed.

- Destroy media according to (NIST) Guidelines for Information Media Sanitization (https://csrc.nist.gov/publications/detail/itl-bulletin/2015/02/nist-special-publication-800-88-revision-1-guidelines-for-media/final).

Additional Guidance for Cloud Computing

Institutions that wish to use cloud computing must work with their cloud service provider to devise an appropriate security plan that meets the general dbGaP Information Security Best Practices as well as these additional requirements that derive from the nature of multi-tenant clouds with default access to the internet. Please refer to the specific cloud service provider for methods, processes and procedures for working with controlled-access data subject to the GDS Policy in the cloud.

General Cloud Computing Guidelines

- Whenever possible, use end-to-end encryption for network traffic. For example, use Hypertext Transfer Protocol (HTTPS) sessions between you and your instance. Ensure that your service uses only valid and up-to-date certificates.

- Encrypt data at rest with a user's own keys. SRA-toolkit includes this feature; other software providers offer tools to meet this requirement.

- Use security groups and firewalls to control inbound traffic access to your instance. Ensure that your security profile is configured to allow access only to the minimum set of ports required to provide necessary functionality for your services and limit access to specific networks or hosts. In addition, allow administrative access only to the minimum set of ports and source IP address ranges necessary.

- Be aware of the top 10 vulnerabilities for web applications and build your applications accordingly. To learn more, visit Open Web Application Security Project (OWASP) - Top 10 Web Application Security Risks. When new Internet vulnerabilities are discovered, promptly update any web applications included in your Virtual Machine (VM) images. Examples of resources that include this information are SecurityFocus and the NIST National Vulnerability Database.

- Review the Access Control Lists (ACLs), permissions, and security perimeter to ensure consistent definition.

Audit and Accountability

- Ensure that data is accessible only to those approved for access, and controls for changing that access are retained by the investigator who submitted the DAR and the appropriate IT staff. A mechanism for monitoring and notification needs to be in place to monitor changes in permission changes.

- Ensure that account access is logged along with access controls and file access and this information is reviewed by the investigator on regular basis to ensure continued secure access.

Image Specific Security

- Ensure images do not contain any known vulnerabilities, malware, or viruses. A number of tools are available for scanning the software, such as Chkrootkit, rkhunter, OpenVAS and Nessus.

- Ensure that Linux-based Images lock/disable root login and allow only sudo access. Additionally, root password must not be null or blank.

- Ensure that images allow end-users with OS-level administration capabilities to allow for compliance requirements, vulnerability updates, and log file access. For Linux-based Images, this is normally through SSH, and for Windows-based virtual machine images, this is normally through RDP.

Best Practices for Specific Cloud Service Providers:

Examples of cloud service provider best practices are provided in the links below, links to the best practices of additional cloud service providers will be periodically appended to this document when they become available. Please be aware that these are provided for convenience only, and do not imply endorsement by the NIH or the United States Government for any of these services, nor does the government guarantee that these links lead to the most current version of these best practices. NIH recommends that investigators consult with their cloud service provider to ensure that they are using the most up to date best practice documents.

Amazon Web Services:

- http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AMIs.html

- http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-best-practices.html

- http://aws.amazon.com/documentation/ec2/

Google Cloud Platforms:

- https://cloud.google.com/developers/articles/best-practices-for-configuring-permissionson-gcp

Others Sources of Information for Cloud Best Practices:

Examples of cloud best practices from organizations that leverage the cloud are provided in the link below. Links to additional documentation will be periodically appended to this document when they become available. Please be aware that these are provided for convenience only, and do not imply endorsement by the NIH or the United States Government for any of these services, nor does the government guarantee that these links lead to the most current version of these best practices. NIH recommends that investigators consult with these organizations to ensure that they are using the most up to date best practice documents.

DNAnexus: https://dnanexus.com/papers/Compliance_White_Paper.pdf

Additional Resources for Testing and Best Practices

Center for Internet Security (CIS)

CIS (http://www.cisecurity.org/) is the only distributor of consensus best practice standards for security configuration. The Benchmarks are widely accepted by U.S. government agencies for Federal Security Information Act (FISMA) compliance, and by auditors for compliance with the International Organization for Standardization (ISO) standard as well as the Gramm-Leach-Bliley (GLB) Act, SarbanesOxley (SOX) Act, federal Health Insurance Portability and Accountability Act (HIPAA), Family Educational Rights and Privacy Act (FERPA) and other regulatory requirements for information security. End user organizations that build their configuration policies based on the consensus benchmarks cannot acquire them elsewhere. See Appendix A for checklists based on CIS best practices, customized for use with controlled-access data. Content of this document has been adapted from NIH Center for Information Technology (CIT), NIST and CIS.

National Institute of Standards and Technology (NIST)

NIST, an agency of the US Department of Commerce provides information security standards and best practices for the federal government. The NIST Special Publications (SP) and Federal Information Processing Standards (FIPS) provide useful and concrete guidance to users of information technology systems (http://csrc.nist.gov/publications/).

United States Government Configuration Baseline (USGCB)

USGCB (http://usgcb.nist.gov) provides security configuration baselines for information technology products widely used across the federal government including desktop computers.

Genomic Data User Code of Conduct

Under the National Institutes of Health (NIH) Genomic Data Sharing Policy, the Genomic Data User Code of Conduct sets forth principles for responsible management and use of large-scale genomic data and associated phenotypic data accessed through controlled access to NIH-designated data repositories (e.g., the database of Genotypes and Phenotypes (dbGaP), repositories established as NIH Trusted Partners). Failure to abide by any term within this Code of Conduct may result in revocation of approved access to datasets obtained through these repositories. Investigators who are approved by NIH to access data agree to:

1. Use datasets solely in connection with the research project described in the approved Data Access Request for each dataset;

2. Make no attempt to identify or contact individual participants or groups from whom data were collected, or generate information that could allow participants' identities to be readily ascertained, without appropriate approvals from the submitting institutions;

3. Maintain the confidentiality of the data and not distribute them to any entity or individual beyond those specified in the approved Data Access Request;

4. Adhere to the NIH Security Best Practices for Controlled-Access Data Subject to the NIH Genomic Data Sharing Policy and ensure that only approved users can gain access to data files;

5. Acknowledge the Intellectual Property terms as specified in the Data Use Certification Agreement;

6. Provide appropriate acknowledgement in any dissemination of research findings including the investigator(s) who generated the data, the funding source, accession numbers of the dataset, and the data repository from which the data were accessed; and,

7. Report any inadvertent data release, breach of data security, or other data management incidents in accordance with the terms specified in the Data Use Certification Agreement.

06/11/2019